Supporting content evaluation of student summaries by Idea Unit embedding

Marcello Gecchele[†] **Hiroaki Yamada**[‡] **Takenobu Tokunaga**[‡] **Yasuyo Sawaki**[◇]

[†]University of Trento
Trento, Italy

[‡]School of Computing
Tokyo Institute of Technology
Tokyo, Japan

[◇]School of Education
Waseda University
Tokyo, Japan

marcello.gecchele@pm.me yamada.h.ax@m.titech.ac.jp take@c.titech.ac.jp ysawaki@waseda.jp

Abstract

This paper discusses the computer-assisted content evaluation of summaries. We propose a method to make a correspondence between the segments of the source text and its summary. As a unit of the segment, we adopt "Idea Unit (IU)" which is proposed in Applied Linguistics. Introducing IUs enables us to make a correspondence even for the sentences that contain multiple ideas. The IU correspondence is made based on the similarity between vector representations of IU. An evaluation experiment with two source texts and 20 summaries showed that the proposed method is more robust against rephrased expressions than the conventional ROUGE-based baselines. Also, the proposed method outperformed the baselines in recall. We implemented the proposed method in a GUI tool "Segment Matcher" that aids teachers to establish a link between corresponding IUs across the summary and source text.

1 Introduction

Summary writing is a complex task involving various linguistic operations, and as such it is useful for developing student linguistic proficiency including text comprehension and composition (Graham and Perin, 2007). The quality of a summary is a good indicator of language proficiency. Therefore, teachers can use summaries to evaluate a student's proficiency. To evaluate the quality of a summary, a teacher has to assess if the summary conveys the important ideas of the source text, as well as the grammatical and lexical correctness. However, finding the corresponding information between the summary and source text is not an easy task for humans.

Another important aspect of summarization is rephrasing. This practice is encouraged as it is a core skill to master, especially for scholars, to avoid plagiarism (Keck, 2014). Rephrasing, however, often obfuscates the bonds between the contents of the source and summary texts and it represents one of the reasons why summary evaluation is such a complex activity.

This paper proposes a support tool for evaluating student summaries in terms of their contents by suggesting the links between the ideas of a source text and its summary. We divide texts into Idea Units (IUs) in order to deal with complex sentences that convey multiple ideas. The IU is defined as a minimal fragment of a text that conveys an "idea" or "thought" coherently (Kroll, 1977). We make correspondence between IUs instead of sentences across the source text and its summary. To circumvent inaccurate IU pairing due to rephrasing we adopt word embedding for the calculation of IU similarity.

2 Related Work

Evaluation is one of the important aspects of the automated text summarization research (Lin and Hovy, 2003). BLEU (Papineni et al., 2002) delivers a similarity score by analyzing n-grams that appear both in the source and summary texts in terms of precision. ROUGE (Lin, 2004) expands on BLEU by providing recall-oriented statistics with n-grams and Longest Common Subsequence. As these measures are based on string matching of n-grams, they fail in making a correspondence between rephrased expressions.

The Pyramid approach (Passonneau, 2009) divides the texts into text fragments named Summary Content Units (SCU). Assuming a set of summaries for a source text, SCUs are weighted based on their frequency over the summary set. The rationale is that frequent SCUs contain important ideas. The score of a summary is calculated by summing up the weight of every SCU in

Proceedings of the Fourteenth Workshop on Innovative Use of NLP for Building Educational Applications, pages 343–348
Florence, Italy, August 2, 2019.

the summary.

Automatic summary evaluation tools based on the Pyramid approach, such as PEAK (Yang et al., 2016) and PyrEval (Gao et al., 2018), are not suitable in educational environments as we cannot expect a number of reliable summaries large enough to certify a proper weighting. In addition, the quality of summaries is not guaranteed due to insufficient student proficiency in comprehension or composition. Their summaries might overlook obscure yet paramount information. These facts lead to imprecise SCU weighting. Lastly, writing a Gold Standard summary is a time-consuming task; therefore we are forced to compare the summaries against the source text directly.

FRESA (Torres-Moreno et al., 2010) is a framework for the evaluation of summaries that relies on the Jensen-Shannon divergence between n-gram probabilities. It scores summaries directly against the source text without reference summaries. A high correlation was reported between the Jensen-Shannon divergence against the source text and the ROUGE or Pyramid-based scores, which are based on the reference summaries. However, as the metric relies on n-grams, such high correlation cannot be guaranteed when summaries use a lot of rephrasing.

3 Segmentation

We divide the summaries and source text into Idea Unit (IU) and make a correspondence between them. The reason why segmentation is necessary can be found in Keck (2014). In Keck's study, the level of rephrasing of student summaries was manually graded by matching sentences that shared some words. This implies that rephrased sentences in the summary borrow at least one term from their source text. Keck mentions that Gist statements were particularly difficult to analyze as they expressed the information described in multiple sentences in a few words. Such constructs are desirable, as they are an indication of an advanced understanding of the language, but finding the corresponding sentences in the source text is difficult. Shorter units than sentences would be more versatile for making a correspondence between the summaries and source text.

Foster et al. (2000) analyzed several segmentation units from the viewpoint of intonation, syntax and semantics. For our purpose, we consider three kinds of syntactic units: IU, T-Unit (Hunt, 1965, 1966, 1970) and C-Unit (Loban, 1963). Despite being a popular approach, the T-Unit is too generous as it includes subordinate clauses in a single unit. Furthermore, the T-Unit is purely a syntactic unit, while IUs and C-Units also serve as a semantic unit. Despite being readopted by multiple scholars over the years, the C-Unit is rather vague in its definition and still retains the T-Unit feature of allowing multiple clauses in a unit. On the other hand, IUs tend to be shorter in length. For instance, it separates relative clauses in different units (Figure 1). Moreover, its rather strict definition suggests a smooth transition into an automatic segmentation algorithm in the future.

C-unit concerns the identification of units. The T-unit and C-unit use orthographic sentences as the unit of analysis. However, identifying orthographic sentences could be a problem in analyzing student summaries, particularly those written by second language learners, due to grammatical errors and punctuation.

In Applied Linguistics, IUs have been employed for in-depth analyses of the content of student summaries in the second language learning and assessment literature (Johns and Mayes, 1990). Accordingly, adopting the IU enables us to interpret our study results in reference to such previous investigations of summary content.

4 Ranking Method

To link two corresponding IUs across the summary and source text, we calculate the similarity between the units based on word embedding. A vector representing an IU is constructed by averaging the vector representation of the words appearing in the unit. We use the GloVe word vectors (Pennington et al., 2014) that have been pretrained with the Wikipedia + Gigaword data. We ignored the words that are not included in the word vector model when constructing the IU vector. We call an IU in a summary "Summary IU" and one in the source text "Source IU" hereafter. Given a Summary IU, its cosine similarity to every Source IU is calculated to create a ranking list of Source IUs that are arranged in descending order of similarity. We called this list "Prediction Ranking".

As a baseline, we use ROUGE-1, ROUGE-2 and ROUGE-L-based rankings. We selected ROUGE as it has proven to be effective in evaluating short summaries of single documents (Lin, 2004).

1. a subject and verb counted as one idea unit together with (when present) a (a) direct object, (b) prepositional phrase, (c) adverbial element, (d) mark of subordination, *or (e) a combination of the above*
2. full relative clauses counted as one idea unit when the relative pronoun was present
   (a) *phrases that are set off by a complementizer are counted as an Idea Unit*
   (b) *subordinate conjunctions and relative pronouns are always attached to the subordinate clause*
3. phrases which occurred in sentence initial position followed by a comma or which were set off from the sentence with commas were counted as separate idea units
   (a) *adverbial conjunctions (e.g.: "However,") are not to be split into separate Idea Units*
   (b) *citations are counted as separeted idea units only when they are set off from the sentence in their entirety*
4. verbs whose structure requires or allows a verbal element as object were counted with both verbal elements as one idea unit
5. reduced clauses in which a subordinator was followed by a non-finite verb element were counted as one idea unit
6. post-nominal -ing phrases used as modifiers counted as one idea unit
7. other types of elements counted as idea units were (a) absolutes, (b) appositives, and (c) verbals
8. *An idea unit can be discontinuous*

Figure 1: Extended definition of IU based on Kroll (1977). Our edits are presented in *italics*.

5 Evaluation

5.1 Data set

Our data set is comprised of two source texts and ten student summaries for each. The sources were taken from the questions in the comprehension section of the IELTS English proficiency test and their topic is "the preservation of endangered languages" and "the impact of noise on cognitive abilities". The summaries were composed by ten Ph.D. students of the University of Cambridge. They were instructed to summarize each source text to about one-fourth of the original length in 15 minutes while maintaining every piece of information they deemed necessary to the correct understanding of the source text. Table 1 illustrates the stats of the data set. The column "Summary" shows the averaged figures of ten summaries.

	Source 1	Summary	Source 2	Summary
Words	996	185.5	807	204.5
IUs	111	20.6	89	24.6
Links	—	18.0	—	21.3

Table 1: Statistics of data set

We manually segmented all texts into IUs according to an extended version of Kroll (1977)'s specification. Our version includes some addenda to define an IU as strictly and as clearly as possible (Figure 1). The extended parts are italicized in Figure 1. Syntactical and grammatical corrections were deemed out of scope and as such the texts were left unedited.

We also manually aligned the corresponding Summary IUs and Source IUs to make a set of correct IU links, pairs consisting of a Summary IU and a corresponding Source IU. No link was assigned to a Summary IU in cases where its content contradicts the source or was entirely fabricated by the student. Our data set includes such linkless IUs since the number of links is less than that of Summary IUs as shown in Table 1. A Summary IU can have multiple links to Source IUs as long as it contains information from those Source IUs. These gold IU links were used in the evaluation of our ranking method.

5.2 Evaluation Metric

For each IU in our set of summaries, we calculated a Prediction Ranking based on four ranking methods: the proposed Vector-based ranking and three ROUGE-based baselines (ROUGE-1, ROUGE-2 and ROUGE-L). We then studied the precision and recall of these rankings to evaluate the effectiveness of our Vector-based model.

The recall and precision are calculated as follows.

$$Precision^{(n)}(s) = \frac{|PR^{(n)}(s) \cap GL(s)|}{|PR^{(n)}(s)|}, \quad (1)$$

$$Recall^{(n)}(s) = \frac{|PR^{(n)}(s) \cap GL(s)|}{|GL(s)|}, \quad (2)$$

where s is a summary, $PR^{(n)}(s)$ is the Prediction Ranking sliced at the top n links for summary s and $GL(s)$ is the set of Gold links for summary s.

We further averaged recall and precision values over all summaries. Figure 2 and Figure 3 show the averaged precision and the averaged recall against the rank threshold n.

5.3 Results

Figure 2 and Figure 3 indicate that our Vector-based method shows comparable performance to the three ROUGE-based baselines in terms of precision but our method outperforms the others in recall. The difference in recall becomes larger according to the increase of the rank threshold n. This result is promising, as the final decision on

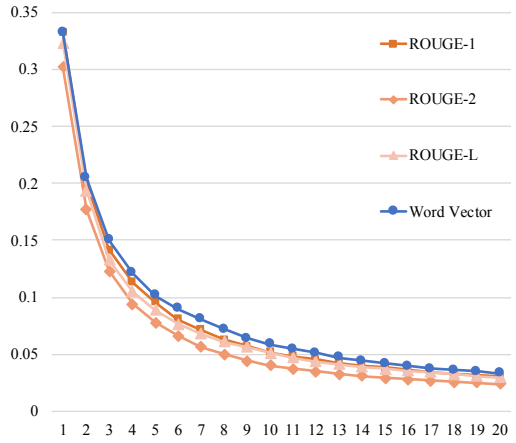

Figure 2: Averaged precision at at rank threshold n

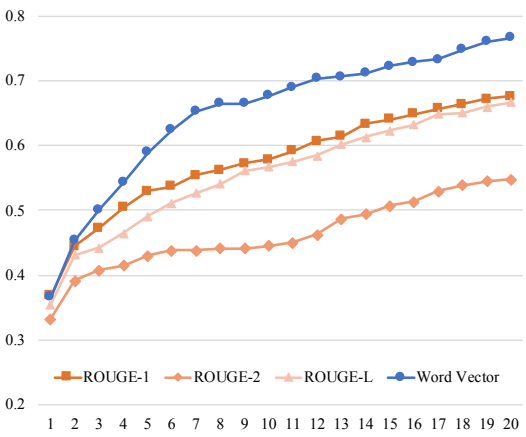

Figure 3: Averaged recall at rank threshold n

what IUs should be linked is left to the end user, i.e. we favor recall.

The recall curve for summary 4B (Figure 4) is an example that shows a particularly big difference between our method and the others. When the ranking threshold is 6, the recall of our model is saturated at 0.82, which is more than the double of the recall of other baselines.

Figure 5 shows an example of the robustness of our method against rephrasing. The Source IU "*that its predictability is more important than how loud it is*" was linked to the Summary IU "*A large factor is not the volume*" by our raters. These two IUs share very few words, but they are close in meaning. Our Vector-based ranking method was able to capture this correspondence while the baselines could not. Indeed, the Vector model ranked the Source IU as the 5th most probable candidate for the Summary IU, while ROUGE-1 and ROUGE-L ranked it as 27th and 26th respectably. ROUGE-2 failed altogether to match the two segments. This data is shown along with the relative

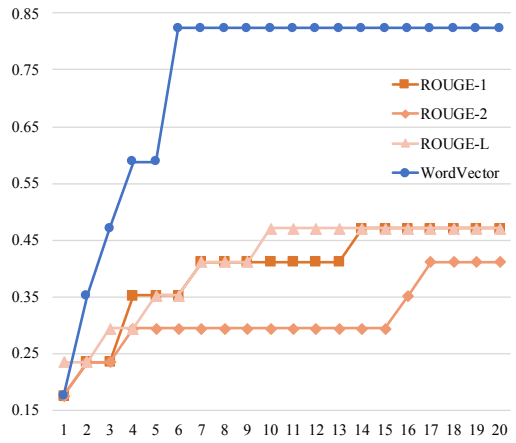

Figure 4: Recall at rank threshold n for summary 4B

Summary 4B

... A large factor is not the volume,\but the predictability of a noise,\as those exposed to quiet unpredictable noises performed worse than those listening to a loud predictable one. ...

Source text 2

... Probably the most significant finding from research on noise is \that its predictability is more important than how loud it is.\We are much more able to 'tune out' chronic, background noise, ...

Figure 5: IU samples with rephrasing.

similarities in Table 2.

6 Segment Matcher: A visual helper for Idea Unit alignment

We built a tool named "Segment Matcher" to aid teachers to establish links between Summary IUs and Source IUs through a graphical user interface. The tool consists of a front end developed entirely in JavaScript to ensure platform independence and a back end Python server to calculate the similarity between IU vecors. The back end server was built in Python inside a Docker container[1] for portability reasons.

The front end presents three different modes of use: Match, Edit and Compare. In the Match mode, the user firstly selects a summary file and its source file that adhere to our data format. Each text should presents one IU per line, with discontinuous IUs having a number prefix followed by a control character. These files are uploaded to the server and Prediction Rankings are returned for every Summary IUs. When the segment rankings have been successfully received, Segment Matcher moves to the link editor.

[1] https://www.docker.com/

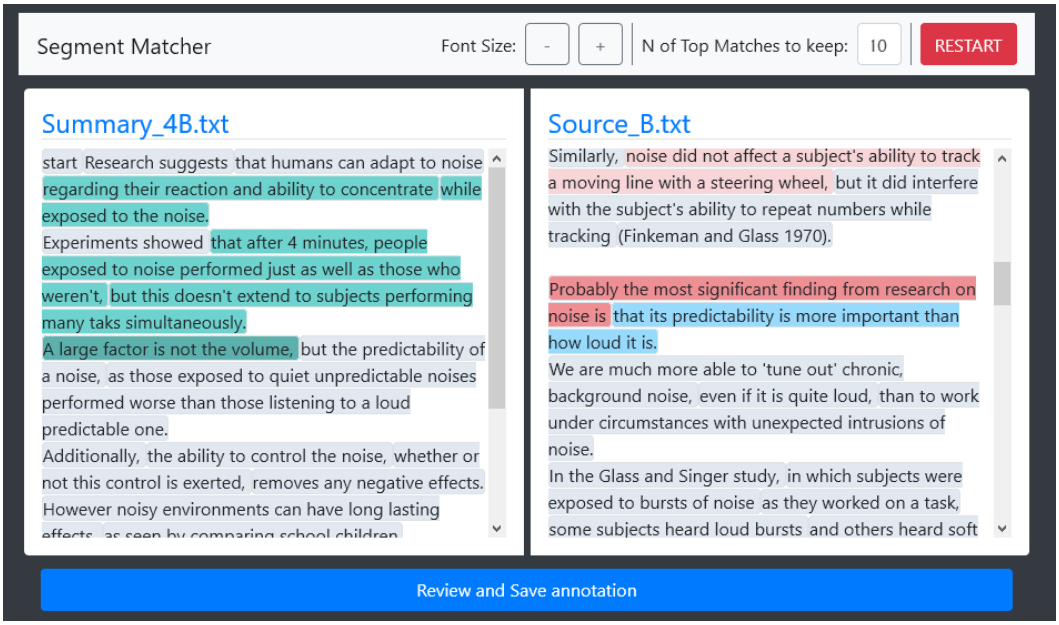

Figure 6: Screenshot of Segment Matcher

	Vector	R-1	R-2	R-L
Rank	5	27	—	26
Similarity	0.91	0.12	0	0.11

Table 2: Rank and similarity for IUs in Figure 5

Figure 6 shows a summary and the source text side by side, with IUs encircled in balloons, signalling that they are clickable elements. The user can link two IUs by first selecting a Summary IU from the left panel and then clicking the relevant Source IU in the right panel. To simplify the user experience, we colored the IU balloons as follows. When a Summary IU is firstly selected, it turns into yellow and the top N most likely candidates for the Summary IU are colored in a different shade of red, from the darkest indicating the most likely candidate to the lightest indicating the least likely one. When the user clicks a Source IU to be linked, the current Summary IU turns into dark green and the linked Source IU turns into blue as shown in Figure 6. The already linked Summary IUs are indicated in light green. The user can choose how many candidates to highlight, with the default being five.

Once the user is satisfied with their work, they can review the alignment by listing the IU links and save the alignment in a CSV file. This CSV file can be modified later via the Edit mode.

The Compare mode allows users to compare the alignments of two different raters, where two alignment CSV files can be selected along with their source texts to show the IU links side by side.

7 Conclusion and Future Work

In this paper, we introduced the Idea Unit (IU) (Kroll, 1977) for the content evaluation of student summaries and proposed a method for aligning IUs across a source text and its summaries. Our aligning method adopts the word embedding technique to deal with rephrased expressions. The experiment with 20 summaries for two source texts confirmed that our proposed method is more robust against rephrasing than the ROUGE-based baselines. The experiment also showed that our method outperformed the baselines in recall. The high recall is favorable as the final decision on the IU alignment is left to the end user.

Adopting the proposed aligning method, we built "Segment Matcher" to aid teachers to establish links between the IUs in a summary and the source text through a graphical user interface. We believe our tool contributes to making the content evaluation of student summaries by teachers more efficient.

In the future, we plan to further improve our work by implementing an automatic segmentation algorithm. This will allow teachers to evaluate summaries without having to segment them into IUs beforehand. We believe this to be a mission-critical feature that has to be implemented before the tool can be considered complete. We also plan to conduct tests in real world scenarios before releasing our tool to the public.

References

Pauline Foster, Alan Tonkyn, and Gillian Wigglesworth. 2000. Measuring spoken language: A unit for all reasons. *Applied linguistics*, 21(3):354–375.

Yanjun Gao, Andrew Warner, and Rebecca Passonneau. 2018. Pyreval: An automated method for summary content analysis. In *Proceedings of the 11th Language Resources and Evaluation Conference*, Miyazaki, Japan. European Language Resource Association.

Steve Graham and Dolores Perin. 2007. A meta-analysis of writing instruction for adolescent students. *Journal of Educational Psychology - J EDUC PSYCHOL*, 99.

Kellogg W Hunt. 1965. *Grammatical structures written at three grade levels*. 3. National Council of Teachers of English Champaign, IL.

Kellogg W Hunt. 1966. Recent measures in syntactic development. *Elementary English*, 43(7):732–739.

Kellogg W Hunt. 1970. Syntactic maturity in schoolchildren and adults. *Monographs of the society for research in child development*, 35(1):iii–67.

Ann M Johns and Patricia Mayes. 1990. An analysis of summary protocols of university esl students. *Applied linguistics*, 11(3):253–271.

Casey Keck. 2014. Copying, paraphrasing, and academic writing development: A re-examination of l1 and l2 summarization practices. *Journal of Second Language Writing*, 25:4 – 22.

Barbara Kroll. 1977. Combining ideas in written and spoken english: a look at subordination and coordination. In Elinor Ochs and Tina Bennett-Kastor, editors, *Discourse across time and space*, volume 5 of *S.C.O.P.I.L.* Los Angeles, Calif.: Dept. of Linguistics, University of Southern California.

Chin-Yew Lin. 2004. Rouge: A package for automatic evaluation of summaries. In *Text Summarization Branches Out: Proceedings of the ACL-04 Workshop*, pages 74–81, Barcelona, Spain. Association for Computational Linguistics.

Chin-Yew Lin and Eduard Hovy. 2003. Automatic evaluation of summaries using n-gram co-occurrence statistics. In *Proceedings of the 2003 Human Language Technology Conference of the North American Chapter of the Association for Computational Linguistics*, pages 150–157.

Walter Loban. 1963. The language of elementary school children.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of 40th Annual Meeting of the Association for Computational Linguistics*, pages 311–318, Philadelphia, Pennsylvania, USA. Association for Computational Linguistics.

Rebecca J. Passonneau. 2009. Formal and functional assessment of the pyramid method for summary content evaluation. *Natural Language Engineering*, 16(2):107–131.

Jeffrey Pennington, Richard Socher, and Christopher D. Manning. 2014. Glove: Global vectors for word representation. In *Empirical Methods in Natural Language Processing (EMNLP)*, pages 1532–1543.

Juan-Manuel Torres-Moreno, Horacio Saggion, Iria da Cunha, Eric SanJuan, and Patricia Velázquez-Morales. 2010. Summary evaluation with and without references. *Polibits*, (42):13–20.

Qian Yang, Rebecca J Passonneau, and Gerard De Melo. 2016. Peak: Pyramid evaluation via automated knowledge extraction. In *Thirtieth AAAI Conference on Artificial Intelligence*.